González-Díaz, I., Birinci, M., Díaz-de-María, F., Delp, E.J. (2017). Neighborhood Matching for Image Retrieval. *IEEE Transactions on Multimedia*, 19(3), pp. 544-558.

DOI: https://doi.org/10.1109/TMM.2016.2616298

Neighborhood Matching For Image Retrieval

Iván González-Díaz, *Member, IEEE,* Murat Birinci, Fernando Díaz-de-María, *Member, IEEE,* and Edward J. Delp, *Life Fellow, IEEE*

Abstract

In the last few years large-scale image retrieval has attracted a lot of attention from the multimedia community. Usual approaches addressing this task first generate an initial ranking of the reference images using fast approximations that do not take into consideration the spatial arrangement of local features in the image (e.g. the Bag-of-Words paradigm). The top positions of the rankings are then re-estimated with verification methods that deal with more complex information, such as the geometric layout of the image. This verification step allows pruning of many false positives at the expense of an increase in the computational complexity, which may prevent its application to large-scale retrieval problems. This paper describes a geometric method known as *Neighborhood Matching* (NM), which revisits the keypoint matching process by considering a neighborhood around each keypoint and improves the efficiency of a geometric verification step in the image search system. Multiple strategies are proposed and compared to incorporate NM into a large-scale image retrieval framework. A detailed analysis and comparison of these strategies and baseline methods have been investigated. The experiments show that the proposed method not only improves the computational efficiency, but also increases the retrieval performance and outperforms state-of-the-art methods in standard datasets, such as the *Oxford 5k and 105k datasets*, for which the spatial verification step has a significant impact on the system performance.

Index Terms

image retrieval, geometric verification, neighborhood matching, robust estimation.

I. González-Díaz and F. Díaz-de-María are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, 28045, Spain, e-mail: {igonzalez,fdiaz}@tsc.uc3m.es. Murat Birinci is with the Department of Signal Processing at Tampere University of Technology, Korkeakoulunkatu 7, 33720 Tampere, Finland. Edward J. Delp is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: ace@ecn.purdue.edu).

This work has been partially supported by the National Grant TEC2014-53390-P of the Spanish Ministry of Economy and Competitiveness.

I. Introduction

Following recent developments in camera and internet technology, size and diversity of image collections keep increasing at an astonishing speed. In particular, *social networks* play a very relevant role in making these images publicly available. *Instagram*, one of the most popular image sharing platforms, today contains more than 30 billion images with contents varying from passion of jazz to popular protests, and has a growth rate of 70 million images per day [1]. Inevitably, such a growth also raises the problem of convenient and efficient access to this content.

Various systems and search engines are available for easy access and retrieval of relevant multimedia content, most of them rely on textual data associated with the visual contents. Despite being efficient and fairly successful, such systems suffer vastly from the well-known *semantic gap* [2] in addition to being highly noisy and ambiguous. In order to address such problems, content-based image retrieval and, in particular, partial or near-duplicate image search where a query image is used to retrieve images that share certain visual elements have become a key element in many image [3], [4] and video [5] search systems.

Inspired by natural language processing methods, most state-of-the-art image retrieval techniques rely on the Bag of Words (BoW) model [6]. Typically, visual words are used to efficiently encode the visual appearance of salient local features, initially described using high-dimensional descriptors, such as SIFT [7].

It is well known that the quantization process inherent to BoW reduces the discriminative power of local descriptors. This reduction in the performance is, in general, limited by using very large visual vocabularies (up to millions of words) [8] that convert BoW into an efficient approximation of direct matching between descriptors. Since the use of such large vocabularies might penalize computational efficiency, more advanced solutions have been proposed which either associate binary signatures with local descriptors to reduce the size of the vocabularies [9], or even circumvent their use [10]. Binary descriptors can also be used at a higher granularity to obtain more compact image representations thereby reducing the computational complexity of the image matching process, or even improving its matching performance. For example, in [11] each image is projected into a category-space and specific query weights are assigned to each bit of the binary hashes to construct a query-adaptive image search scheme. In [12] binary codes are computed that minimize the quantization error of mapping high dimensional data to the vertices of a zero-centered binary hypercube, demonstrating better results than previous approaches.

Despite of these improvements, the baseline BoW model does not take into consideration the geometric relationships between local features, which strongly limits its performance in large-scale image search systems. For example, non-relevant images might share many local descriptors with the query and lead to false positives. To overcome this limitation, a final re-ranking step is typically performed to improve the quality of the initial ranking either by checking the geometric consistency of the matches [13], [9], [8], or by adding additional information modeling users' preferences or visual attention [14]. Since such re-ranking methods are often computationally expensive and significantly increase the retrieval time, they are commonly used only for a subset of images (the top-ranked images).

This paper focuses on the spatial verification step of a partial or near-duplicate image search system. We build our model on the *neighborhood matching* (NM) method we proposed in [15]. Rather than simply using NM as a filtering stage that eliminates false correspondences between images, we use NM for the geometrical verification step in large-scale image retrieval. The goal of our approach is to demonstrate that the computational complexity of this stage can be significantly reduced for a given performance level. There are two main contributions in this paper:

- We have developed several strategies that incorporate the NM method into the image retrieval framework. They will be compared in terms of both retrieval performance and computational efficiency and we will show that some of them not only reduce the computational complexity at a given depth in the re-ranking step (measured as the number of images being re-ranked), but also improve the performance with respect to the traditional solutions.
- We provide a detailed analysis of the use the NM techniques to the image retrieval problem and describe the most appropriate configurations. We will also demonstrate that our NM-based re-ranking step can be combined with other techniques involved in processing, such as extensions or improvements of the computation of the initial ranking, or query expansion methods working using geometrically-verified images.

The paper is organized as follows. In Section II we discuss the previous work related to large-scale image retrieval and search. In Section III we briefly describe the neighborhood matching method we proposed in [15]. The proposed method for efficient geometric verification and the experimental results are described in Section IV and Section V. Finally, Section VI concludes the paper and introduces our future work.

II. Related Work

Although there been some recent work that uses deep Convolutional Neural Networks (CNNs) to produce alternative multi-scale image representations [16], [17], [18], the Bag of Words (BoW) model has been the “de-facto” approach for image retrieval in recent years. Typically, local visual features are used to represent local image regions identified by several detectors, (such as SIFT [7], SURF [19], Harris [20], Hessian [21]) and described using well-known local-shape descriptors (such as SIFT [7], RootSIFT [8], SURF [19]). Since the descriptors are usually of high dimensionality, they are clustered into “visual words” in order to achieve more compact and more generic representations. Each image is then represented by a histogram of visual words that model how many of the “words” they accommodate. The histograms are later used to compare different images. Some improvements over the baseline BoW approach have been proposed in the literature. Relevant examples can be found in [9]-[22], where the authors encode the location of the descriptors within the Voronoi-cells associated with their visual words as a Hamming-Embedding. This idea is later improved in more recent works, such as [23], where the Cartesian product used in the nearest neighbor search is decomposed in lower dimensional subspaces, each subspace is then quantized separately. In [24], where observations about the matching between descriptors are incorporated to the embedding method. In [25], multiple soft-assignment of descriptors to various words in the vocabulary are examined. The modeling of the burstiness phenomenon (repetitive elements in an image) was introduced in[26].

The histograms of word occurrences or even the individual matches between visual descriptors of two images usually do not take into account the spatial information of the underlying local features. Since visual words are extracted from local patches, it is rather easy to match (judge as similar) totally different images by only checking how many features are present in both images. The importance of spatial information is discussed in detail in [15] where the prägnanz from Gestalt psychology are taken as reference to impose spatial constraints during feature matching. The simplest and early approaches incorporating information about feature spatial distribution along the images can be found in [27], where the authors proposed matching localized subimages. Also in [28] a pyramid-like representation are used for incorporating spatial distribution where images are divided into a finer grid at each pyramid level and histograms are calculated from each grid cell and then concatenated to form the Spatial Pyramid Matching (SPM). More recent efforts have also been made in order to improve SPM [29], [30], [31]. Graph theory has also been used in order to extract spatial information [32], [33], [34]. For example, in [29], Ren et al. partitioned the image into a predefined number of graphs, and a BoW approach was then used for each sub-graph independently to finally represent an image as a set of BoWs (which they called bag-of-bag-of-words).

Alternatively, spatial information has been exploited as a post-processing stage via geometrical verification so that retrieved images are re-ranked based on the verification. Two approaches are the main trends in geometric verification: methods that explicitly compute the geometric transformation between images that show the same object/scene, and methods that implicitly verify the geometric consistency of matches without explicitly computing the transformation. Among the former, in [35] the authors proposed a generative probabilistic framework that concurrently models the global geometric transformation between matched images and spatial location of matched objects. During the last few years RANSAC (Random Sample Consensus) [36] has become the most prevalent method to verify whether the spatial distribution of features in two images match. RANSAC is a robust iterative method for the estimating a model from a set of observations under the assumption that it may possibly contain outliers. The method is popularly used in computer vision applications, such as object detection/recognition and camera pose estimation for estimating the geometric transformation between two images using point-wise matches of local image features. RANSAC's randomness and large computational cost either prevents its application on large-scale problems or limits it to a small set of top-ranked images. There have several approaches proposed to alleviate computational problems of using RANSAC [37][38][8].

With respect to the second set of verification methods described above, the goal being reducing the computational complexity, other studies have described use of the Hough transformation space. Jegou et al. proposed the use of "weak geometry consistency (WGC)" to filter matching descriptors that are not consistent in terms of scale and angle in Hough space [9]. Other works, such as [39][40], also examined the consistency of matches. Wu et al. [41] suggested to bundle regions and point features together in order to increase discriminating power. It should be noted that although such bundling may recall the grouping of point features we proposed in [15], it requires that two separate features be extracted from every image. The work in [42] also extended the study of individual correspondences and considered pairwise geometric relations between matches to improve the verification accuracy.

The goal of this paper is to merge both families of methods into a unified framework. We will demonstrate this by grouping local features and matches into spatial neighborhoods and show it is possible to dramatically decrease the computational complexity of a model-based geometric verification step while maintaining or even improving the retrieval performance.

The study of geometric relationships between matches is not new in the literature and has been extensively studied using BoW. In particular, in [43] the authors introduced the concept of *visual phrases*, groups of visual words that explicitly encode their consistent co-occurrence in many images. The same idea is explored using several geometric configurations: in [40] translations between matching points are encoded and grouped among co-occurring matches, while in [44], radial relations between pair of matches are described. Furthermore, in [45], the phenomena of polysemy and synonymy are modeled by analyzing the co-occurrence of visual words in all images and in groups of images belonging to a particular semantic category.

Our proposed approach differs from the above in several aspects: first, neighborhoods are built around very reliable matches that we will call strong matches. Although other matches (weak and strong) may be identified in the surroundings of a strong match to validate the presence of a neighborhood, their spatial relationships are not explicitly encoded. This will limit the computational complexity. Indeed, the location and the cardinality of a neighborhood (the number of matches that belong to the neighborhood) are the only metrics we use. Hence, our approach has to be understood more like a pre-filtering stage that allows removing false matches than a way to encode complex spatial relationships between pairs or even groups of words. This paper aims to incorporate the concept of neighborhoods to the geometric verification step of a retrieval system, rather than encoding this information to generate more complex image signatures. Finally, it is also worth mentioning that the generation of neighborhoods is completely category-agnostic and does not require any previous analysis of the dataset.

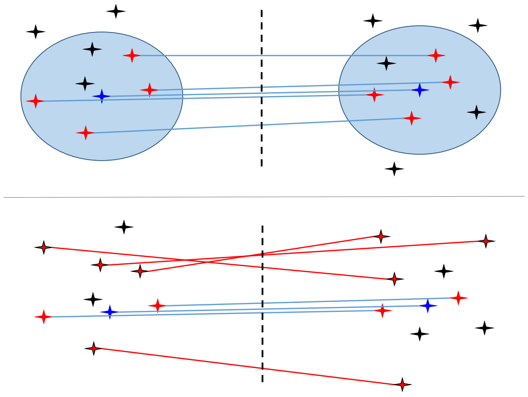


Fig. 1: Illustration of Neighborhood Matching (top) compared to regular matching (bottom). The blue points indicate a strong match and become the center of a neighborhood, which is then completed by other weak matches (in red). Isolated matches that do not belong to any neighborhood are discarded (matches that appear in the lower figure and not in the top figure). We also relax the constraints of the distance ratios of matches within a neighborhood and considering more than one candidate match per point in the query, we can find new matches that would not appear in regular matching (matches that appear in top figure and not in the lower figure). .

III. Neighborhood Matching

We described in [15] that object matching is other than the mere sum of individually matched feature points. We based this postulate on the motto of Gestalt psychology "The whole is other than the sum of its parts". We developed perceptual grouping of feature points - based on spatial proximity - in order to group individually matched features, we called this Neighborhood Matching (NM). In other words when two features are matched, their neighboring features should also match in order to improve the outcome. Fig. 1 illustrates how two neighborhoods in two images are matched using NM compared to regular matching. Initially, the blue feature points are found to be matching. Next, the neighboring feature points around them are determined. Many approaches can be used in order to define what "neighboring" means. Whereas a fixed region around the initial match as in Fig. 1 would be the simplest way, one may also can consider incorporating the scale of the feature point (e.g. SIFT points) while determining the size of the region. Our original approach for NM [15] suggested using nearest neighbor approach where the N spatially closest points to the initial matching point are taken as the "neighboring points". Next, feature points within the neighborhoods are matched with a minimum number of matches required for the entire neighborhood to be considered as matched. Otherwise all matches, including the initial one, are discarded.

Note that if the same similarity threshold is used for both the initial and neighboring matches, NM simply filters out the incorrect matches relative to regular matching. However, if the similarity threshold is loosened for the neighboring matches, it can also increase the total number of matches. Since blindly decreasing the similarity threshold would increase the number of incorrect matches significantly, Lowe suggested using a similarity ratio in order to filter potential incorrect matches [7]. That is, the best and second best matches for a feature point need to be different enough, otherwise it is regarded "ambiguous" and discarded.

In this paper we propose to use NM as a preprocessing step before RANSAC. The potential benefits are twofold: First, RANSAC requires a minimum proportion of inliers to perform properly, NM increases RANSAC's performance by increasing the ratio of inliers. In other words RANSAC is more likely to find the correct transform between matching points, hence better filtering of outliers. Second, as discussed in Section II, RANSAC starts from a random set and iteratively approaches to its final result. If the ratio of outliers is high, it is more likely for that a random set will include outliers and take longer for RANSAC to converge. Hence, increasing the ratio of inliers naturally decreases its convergence time and its computational complexity. Therefore, NM may provide a solution to make it RANSAC more practical.

In order to demonstrate the benefits of NM, we utilized it for the spatial verification step of an image retrieval system. The following section briefly describes the underlying framework and multiple approaches we examined to incorporate NM.

IV. Neighborhood Matching For Image Retrieval

A. Baseline Approach For Image Retrieval

Given a query image I^q and a set of M reference images I^m $\{m = 1, ..., M\}$, the goal of an image retrieval system is to generate a ranking of reference images based on their visual similarity with the query. Following the approach in [8], an image retrieval system computes a baseline ranking using a Bag-of-Words (BoW) model with a very large vocabulary and then a geometric verification step that uses a re-ranking process over the best ranked images is done. These two main phases are briefly reviewed in the next subsections.

1) BoW-based image retrieval: The computation of a visual similarity-based ranking using a Bag-of-Words model involves the following steps. For each image, a set of local regions of interest (keypoints) are detected. Then, the appearance of these keypoints is characterized using descriptors such as SIFT [7]. The descriptors are then vector-quantized using a visual vocabulary (up to 1 million words in our case) in order to generate the image signatures which are normalized to be independent of the number of detected keypoints. The image signatures can be later compared using a similarity metric between the query and the reference images to produce a baseline ranking. For a detailed description of BoW-based image retrieval the interested reader is referred to [8]. It is well known that combining this scenario with the use of very large vocabularies becomes an efficient approximation of the matching process between images descriptors.

2) Geometric verification: The baseline ranking described above is later refined by means of a geometric verification process: given a query and a reference image, an initial set of N_p keypoints in the query and those detected in the reference image, a keypoint matching process is performed which, for each keypoint in the query, finds the most similar keypoint in the reference image according to their respective visual descriptors. In order to filter out potentially false matches, two thresholds are used: 1) an absolute threshold of the distance between matched keypoints ($TH_{abs} = 0.3$ in our experiments); and 2) a relative threshold TH_{ratio} of the ratio of the distances between the query keypoint and the first and second nearest neighbors $r^{12} = \frac{d_1}{d_2}$ [7]. The goal is to remove ambiguous matches for which there are several similar points. As a result of the use of these two thresholds, a set of N_c candidate matches are generated between the query and the reference images.

The geometric verification of the previously determined N_c candidate matches is then complete. The goal is to compute a global geometric transformation between the two images and filter those matches that do not follow the transformation (false matches). In this paper, a 3x3 Homography matrix H is used to geometrically relate matched points.

Considering the general case where the initial set of N_c matches contains both true and false matches, RANSAC (RANdom SAmple Consensus) [36] is usually employed to obtain a robust estimation of the transformation. RANSAC is able to estimate models even in presence of outliers (in our scenario, it estimates the global transformation between images in presence of false matches). The concept here is that RANSAC can find in the entire dataset, with probability p ($p = 0.99$ in our experiments), a sample set of data which is free of outliers. This outlier-free sample will provide a good estimation of the actual geometric transformation between the two images.

Given the size s of the sample set needed to compute a model (in our experiments, we require $s = 4$ pairs of matched points to estimate a homography), and the proportion of inliers in the initial data set $\alpha = \frac{\#inliers}{\#data}$, one can easily estimate the number of iterations needed to find a sample free of outliers:

$$k = \frac{log(1-p)}{log(1-\alpha^s)} \tag{1}$$

Since α is in general unknown, for each iteration:

1) Draw a random sample containing $s = 4$ matches.
2) Estimate the homography H using the selected sample.
3) For each data point in the entire data set, compute the corresponding estimation error given H. If the error is below a threshold ($TH_{RANSAC} = 0.1$ in our experiments), the data point is considered an inlier.
4) When increasing the number of inliers, update the proportion of inliers α and the estimated number of iterations k.

This iterative process finishes when the number of iterations reaches or exceeds k.

In order to estimate a homography between two images, we use the *Direct Linear Transformation (DLT)*. For a detailed description of the method the reader is referred to [46]. In the next paragraphs we will provide a brief summary to introduce the corresponding notation.

We want to find a homography matrix H such that the vectors $\mathbf{x}'_i$ and $H\mathbf{x}_i$ corresponding to a matched pair are parallel and therefore only differ in magnitude by a factor of w'_i. Hence we look for a H matrix such that:

$$\mathbf{x}'_i \times H\mathbf{x}_i = 0 \tag{2}$$

where $\mathbf{x}_i = (x_i, y_i, 1)^T$ and $\mathbf{x}'_i = (x'_i, y'_i, w'_i)^T$. Denoting the j-th row of H as $\mathbf{h}^{jT}$, we can re-write:

$$H\mathbf{x}_i = \begin{pmatrix} \mathbf{h}^{1T}\mathbf{x}_i \\ \mathbf{h}^{2T}\mathbf{x}_i \\ \mathbf{h}^{3T}\mathbf{x}_i \end{pmatrix} \tag{3}$$

And transform the cross-product $\mathbf{x}'_i \times H\mathbf{x}_i$ into a linear matrix-vector product:

$$\begin{bmatrix} w'_i\mathbf{x}_i^T & \mathbf{0}^T & -x'_i\mathbf{x}_i^T \\ \mathbf{0}^T & -w'_i\mathbf{x}_i^T & y'_i\mathbf{x}_i^T \end{bmatrix} \cdot \begin{pmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{pmatrix} = \mathbf{0} \tag{4}$$

where we have removed the last row of the system of equations since it is a linear combination of the first two rows. These equations hold for any value of w'_i. Hence, without loss of generality, we can consider that $\mathbf{x}'_i$ is on the image plane ($w'_i = 1$). Theses equations have the form $A_i\mathbf{h} = \mathbf{0}$, being A_i a 2×9 matrix and $\mathbf{h}$ an 1×9 vector of unknowns.

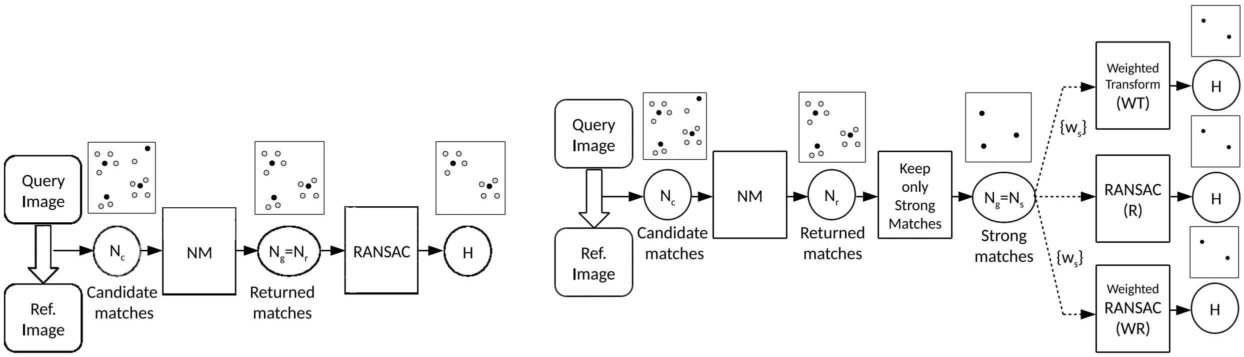
In order to estimate a homography H (through vector $\mathbf{h}$) one can consider a set of s matches (4 or more) to generate a $2s$-equation system: $A\mathbf{h} = \mathbf{0}$. If $s = 4$, the rank of A is 8 and an exact solution can be determined up to a non-zero scale factor. If $s > 4$, the system is over-determined, there is not exact solution, and therefore one should find $\mathbf{h}*$ that minimizes the norm $||A\mathbf{h}||$ instead. In both cases, the constraint $||\mathbf{h}|| = 1$ can be imposed to avoid the trivial solution $\mathbf{h} = \mathbf{0}$ without loss of generality The optimal solution will be given by:

$$\begin{aligned} \mathbf{h}* &= \arg\min_{\mathbf{h}} ||A\mathbf{h}|| \\ & s.t. ||\mathbf{h}|| = 1 \end{aligned} \tag{5}$$

In [46] this constrained optimization problem is solved using Singular Value Decomposition (SVD). In [46] it is noted that this approach becomes unstable in the presence of noise, so that a normalization step is needed to ensure that the solution converges to the correct result. The details about this extension can be found in [46].

B. Neighborhood Matching For Highly Efficient Geometrical Verification

Given the previously described baseline image retrieval system, we suggest using NM mainly to reduce the computational burden associated with geometrical verification. Although RANSAC is a very effective



(a)

(b)

Fig. 2: Processing pipeline of the proposed strategies to incorporate NM. The filled dots represent strong matches whereas unfilled dots represent weak matches. a) NM-based Filtering using RANSAC (NMF+R): NM analyzes the initial set of N_c candidate matches and provides a final set of $N_g = N_r$ matches to the geometric verification step (RANSAC); b) three alternative approaches (SNM+R, SNM+WT, SNM+WR) using only the $N_g = N_s$ strong matches associated with each neighborhood. The strong matches location and the cardinalities of their corresponding neighborhoods are used by the geometric verification in three ways: RANSAC without considering the cardinalities (R), a weighted RANSAC considering the cardinalities (WR), or directly computing the weighted transform without using RANSAC (WT).

technique to deal with significant proportions of false matches, it is also very time consuming. To be more precise, we propose several ways to use NM with the same common purpose: reducing the number of potentially false matches that are given to RANSAC for geometrical verification.

Given a pair of images coming from the top part of the BoW-based similarity ranking, a search for matches between the keypoints of the query image and those of the reference image is done, generating N_c candidate matches. Subsequently, NM is used to validate these matches by requiring that some neighboring keypoints also agree with each candidate match by matching together. As a result, the original number of candidates matches N_c is reduced to a significantly lower number N_s of what we call *strong matches*, one per *matched neighborhood* (when there are not enough matches in the neighborhood, the neighborhood itself does not match and the original candidate match is discarded). The NM process not only returns the strong matches but also all the *weak* matches in each neighborhood around the strong match. We denote the total number of matches returned by the NM as N_r. Furthermore, we associate a cardinality C_s with each strong match s, given by the number of matches that fall within the neighborhood of the strong match.

Below we examine several approaches to take advantage of NM in order to make the geometrical verification more efficient. We denote as N_g the total number of matches passed to the geometrical verification step.

NM-based Filtering + RANSAC (NMF+R): In this case, NM becomes a simple pre-filtering stage that filters potentially false matches before the estimation of the geometric transformation. Hence, only those matches belonging to a matched neighborhood, i.e., a total of $N_g = N_r$ matches, are passed to RANSAC for geometrical verification. This is illustrated in Fig. 2a.

Strong Neighbor Matches + RANSAC (SNM+R): Note that, within a neighborhood, all the weak matches will follow the same or a very similar geometric transformation than that of the strong match (the one originating the neighborhood). In this model we remove all the weak matches and keep only the strong match that generated the neighborhood. This model dramatically reduces the number of points passed to RANSAC for geometrical verification (from N_c to N_s). This is illustrated in Fig. 2b (middle part).

Strong Neighbor Matches + Weighted Transform (SNM+WT): Since the NM process allows us to filter potentially false matches, in this model we estimate the geometrical transformation directly from the strong matches (assuming, therefore, that all of them are inliers), avoiding the use of the iterative and time-consuming RANSAC-based estimation. In order for estimation to be more robust, we propose a reliability measure associated with each strong match that intends to capture the actual strength of each match according to how well the whole neighborhood matches. In particular, we suggest a weighted version of the DLT that takes into account these reliability measures. Starting from the set of ($N_s > 4$) strong matches, and in contrast to Eq. (6), we now minimize a weighted norm $||WA\mathbf{h}||$ where W is a N_sxN_s diagonal matrix $W = \text{diag}\{w_s\}$. Each element in the diagonal contains a reliability measure (weight) w_s associated with its corresponding neighborhood. Specifically, given a strong match s with a cardinality C_s, we have chosen an exponential weighting function of the form:

$$w_s = \sum_{j=1}^{C_s} \exp\left(-\gamma \, r_j^{12}\right) \tag{6}$$

where γ is a free parameter that has been heuristically determined to be $\gamma = 100$. This is illustrated in Fig. 2b (top part).

Strong Neighbor Matches + Weighted RANSAC (SNM+WR): in this last model, we propose to take advantage of the reliability measure associated with each strong match within RANSAC. In particular, instead of simply counting the number of inliers, we propose to choose the best transformation according to a global reliability measure computed over the inlier matches. To be more precise, the proposed *weighted RANSAC* selects the transformation associated to the iteration k^* that maximizes this global reliability measure:

$$k^* = \underset{k}{\operatorname{argmax}} \sum_{s \in I_k} w_s \tag{7}$$

where I_k is the set of inliers obtained at iteration k of RANSAC, and w_s denotes the above described weight associated with the strong match s. This is shown in Fig. 2b (bottom part).

C. Analysis of the Computational Complexity

Since the purpose of using NM for geometrical verification is to reduce computational complexity, we will compare our various approaches with respect to complexity. We have developed a theoretical measure that estimates the computational complexity of the geometrical verification in terms of execution time and corresponds to the execution times we observed in our experiments. Our method is therefore an approximation of the computational complexity, and depends on some execution times associated to basic operations (see in table II) which may differ between architectures.

The total time associated with the geometric verification t_{GV} can be written as:

$$t_{GV} = R\left(t_M + t_{NM} + t_R\right) \tag{8}$$

where R is the number of re-ranked images, i.e., the number of reference images considered for geometric verification (the depth of the re-ranking step); t_M is the time associated with the computation of matches between every pair of images; t_{NM} is the time taken by the NM algorithm; and t_R is the time consumed by the RANSAC-based estimation of the global geometric transformation between two images. Below we describe this in detail.

The time associated with the *computation of matches* between two images t_M is closely related to the strategy used to find the match. In our experiments, we have used the fast approximate nearest neighbor strategy described in [47]. Denoting the average number of detected keypoints per query image by $\bar{N_p}$ and the time needed to compute a nearest neighbor by t_m, we can write t_M as:

$$t_M = \bar{N_p} t_m \tag{9}$$

The time associated with the NM algorithm is mainly that needed to *compute the neighborhoods* and depends on both the number of candidate matches N_c and the time required to compute an individual neighborhood around a strong match t_s. Therefore, t_{NM} can be written as:

$$t_{NM} = \bar{N_c} t_s \tag{10}$$

where, again, $\bar{N}_c$ stands for the average number of candidate matches per query-reference pair.

Finally, the total execution time associated with the *RANSAC geometric verification* can be approximated as follows:

$$t_R = \bar{k}\left(t_h + \bar{N}_g t_e\right) \tag{11}$$

where $\bar{k}$ is the average number of samples drawn from RANSAC, i.e., the number of iterations; t_h is the execution time of generating a model given a random sample, i.e., the execution time to compute a homography transformation H for a given sample; $\bar{N}_g$ is the average number of matches per query-reference pair that is passed to RANSAC ($\bar{N}_g = \bar{N}_r$ for NMF+R and $\bar{N}_g = \bar{N}_s$ for other cases); and t_e is the running time needed to evaluate a transformation on one match.

Putting everything together we obtain the final expression for execution time, t_{GV}:

$$t_{GV} = R\left(\bar{N}_p t_m + \bar{N}_c t_s + \bar{k}\left(t_h + \bar{N}_{gm} t_e\right)\right) \tag{12}$$

According to this expression, when comparing the proposed approach based on neighborhood matching with a baseline geometric verification using RANSAC, we note the following:

1) The term $t_{NM} = \bar{N}_c t_s$ comes from the neighborhood computation (previous to the geometric verification) and, consequently, $t_{NM} = 0$ for the baseline approach.
2) It should be noted that the model SNM+WT avoids the RANSAC-based estimation by using a Weighted Transform on the strong matches resulting from the NM process. Therefore, in this case we remove from the previous expression the execution time associated with RANSAC and set $t_R = 0$.
3) Our proposed approach aims to reduce the number of matches passed to the RANSAC-based geometric estimation $\bar{N}_g$ as well as to improve the proportion of inliers α, thus reducing the number of iterations k needed to ensure a sample free-from-outliers with a given probability p.

For the experimental evaluations, average values for every parameter will be estimated so that approximate execution times can be obtained. Table I is a summary of the main variables used in our approach.

V. Experimental Results

The experiments have been organized into five subsections. First, the datasets and the experimental protocol are described. Second, the parameters of the proposed method which are relevant to the study of the complexity are analyzed. Third, with the purpose of validating our hypotheses, different versions

TABLE I: Summary of the main variables involved in the proposed approach

Variable	Description
N_p	Number of detected keypoints in the query
N_c	Number of candidate matches (true and false)
N_s	Number of strong matches resulting from NM
N_r	Total number of matches resulting from NM
N_g	Number of matches passed to the geometric verification stage
R	Number of re-ranked images by the geometric verification stage
t_{GV}	Total execution time associated with the geometric verification stage
t_M	Total execution time needed to compute matches
t_{NM}	Total execution time needed to compute the neighborhoods
t_R	Total execution time consumed by the RANSAC-based estimation
t_m	Execution time to compute a nearest neighbor (individual match)
t_s	Execution time to compute an individual neighborhood (strong match)
k	Number of iterations taken by RANSAC
t_h	Execution time needed to compute a homography
t_e	Execution time needed to evaluate a homography on a match

of the proposed method are compared with a baseline image retrieval system. Fourth, our NM-based re-ranking step iscombined with other techniques to construct a retrieval system which is compared with state-of-the-art methods. Last, we describe some error analysis to provide better understanding of the limitations of our approach.

A. Datasets and Experimental Setup

We have used three complementary datasets for our experiments:

- The *Oxford 5K dataset* [48]: it contains 5,062 high-resolution images (1024x768) showing either one of the Oxford landmarks (the dataset contains 11 landmarks) or other places in Oxford. The database includes 5 queries for each landmark (55 queries in total), each of them including a bounding box that locates the object of interest. For each query image, this dataset contains hundreds of relevant references, so that the performance will strongly depend on deep positions of the ranked similarity list.
- The *Oxford 105K dataset* [48]: this dataset is a super set of the Oxford 5K dataset in that 100k distracting images downloaded from Flickr have been added. It will allow us to evaluate the performance for a large-scale retrieval system.

TABLE II: Average execution times (ms) by each task for the geometric verification process.

$t_m(ms)$	$t_s(ms)$	$t_h(ms)$	$t_e(ms)$
0.0284	2.608	0.417	0.0131

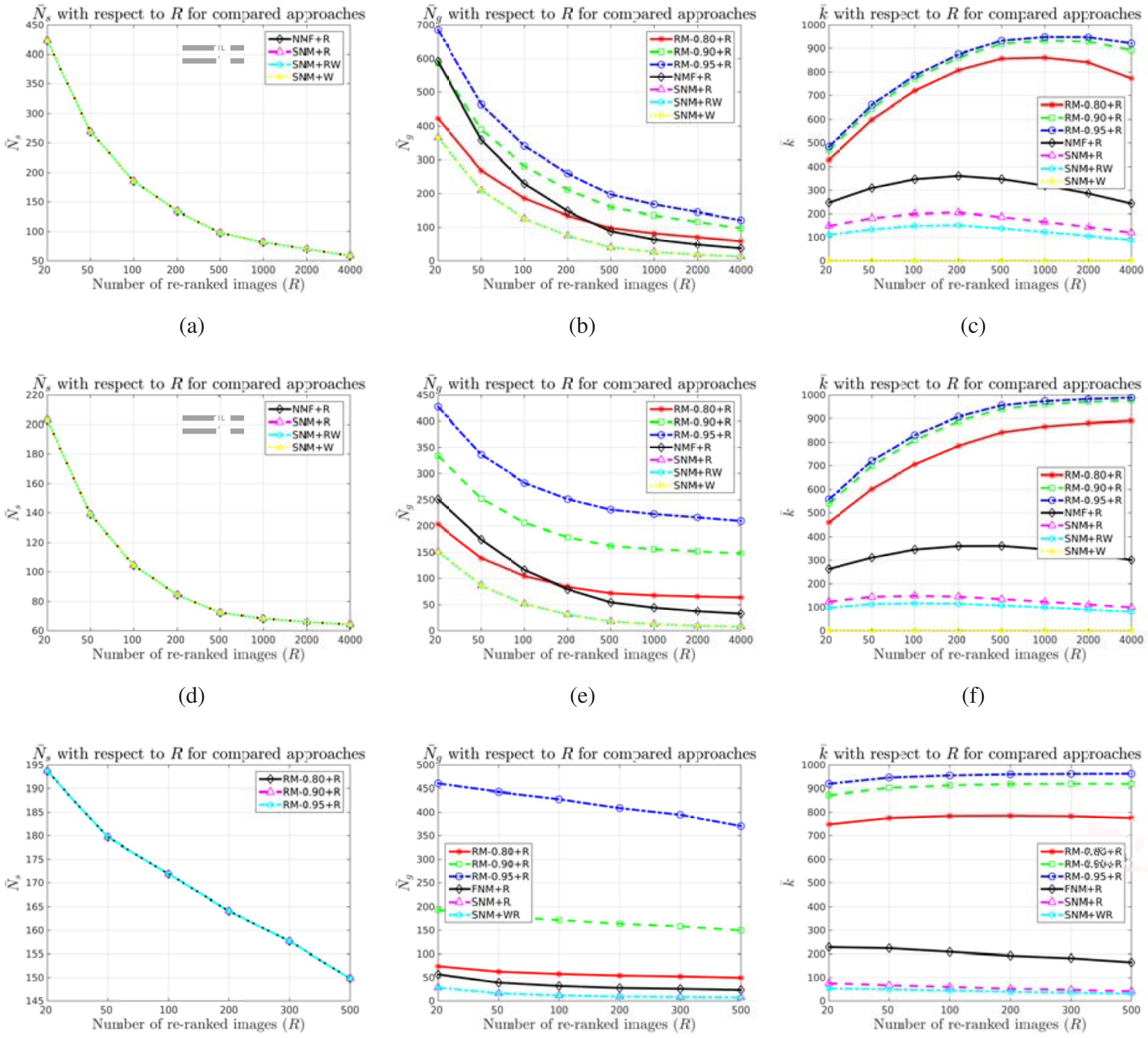
- The *INRIA Holidays dataset* [9]: this is a dataset with 1,491 personal holiday photos, in which several transformations or artifacts can be evaluated: rotations, viewpoint variance, illumination changes, and non-rigid deformations due to moving elements (e.g. clouds). This dataset contains 500 image groups or scenes, with the first image of each group being the query (500 queries). A few images (1-5) are relevant for each query, so that the assessment over this dataset will depend more on the initial positions of the ranked list.

In order to establish a meaningful comparison, we followed the feature extraction protocol described in [8]. In particular, we detected salient points using the affine-invariant Hessian detector [49]. Then, we describe the local region around these keypoints with a 128-dimensional SIFT descriptor [7]. Subsequently, a Bag-of-Words (BoW) model is used; in particular, we employed the same BoW as in [8] with the 1M-sized hard-assigned vocabulary. Finally, we do a re-ranking step using RANSAC [36] starting from what we call *regular matching* (RM), as opposed to *neighborhood matching* (NM), with $TH_{abs} = 0.3$ and three different values for $TH_{ratio} = 0.80, 0.90, 095$. This baseline system is called **RM-TH_{ratio}+R** in the experiments, where TH_{ratio} takes any of the previously mentioned values. It is also worth noting that when Neighborhood Matching is used, it always starts from the regular matches coming from $TH_{ratio} = 0.80$.

We use Average Precision (AP) [50] and mean AP (mAP) for evaluation metrics. mAP is obtained by averaging results for all the queries in each dataset. With respect to execution times in the experiments, all the code has been implemented using MATLAB in except of the features detection and description (to perform a fair comparison. We used the executables suggested by the authors of each datasets [48], [9]), and the optimized MEX-routines (C,C++) for FLANN used for approximate nearest neighbors computation [51]. All the experiments have been executed with single-threading on a computer with 8-cores and a 3GHz processor with 32GB of RAM.

B. Analysis of parameters related to complexity

In order for a fair comparison of the different versions of the geometric verification process proposed in this paper, we compute the mAP achieved by each of the methods for a range of computational complexity



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3: A comparative illustration of the average values of a) $\bar{N}_s$, b) $\bar{N}_g$, and c) $\bar{k}$ as a function of the number of re-ranked images R for all the examined approaches. These parameters are defined in Sec. IV-B and Table I. Top row: Results for Oxford 5k dataset. Middle row: Results for Oxford 105k dataset. Bottom row: Results for Holiday dataset.

levels, measured in execution time (seconds). Two are the main factors that affect the computational complexity: a) the processing pipeline of each geometric verification method, and b) the number of re-ranked images R in the geometric verification.

As described in Section IV-C, in order to obtain a reasonable approximation of the execution time

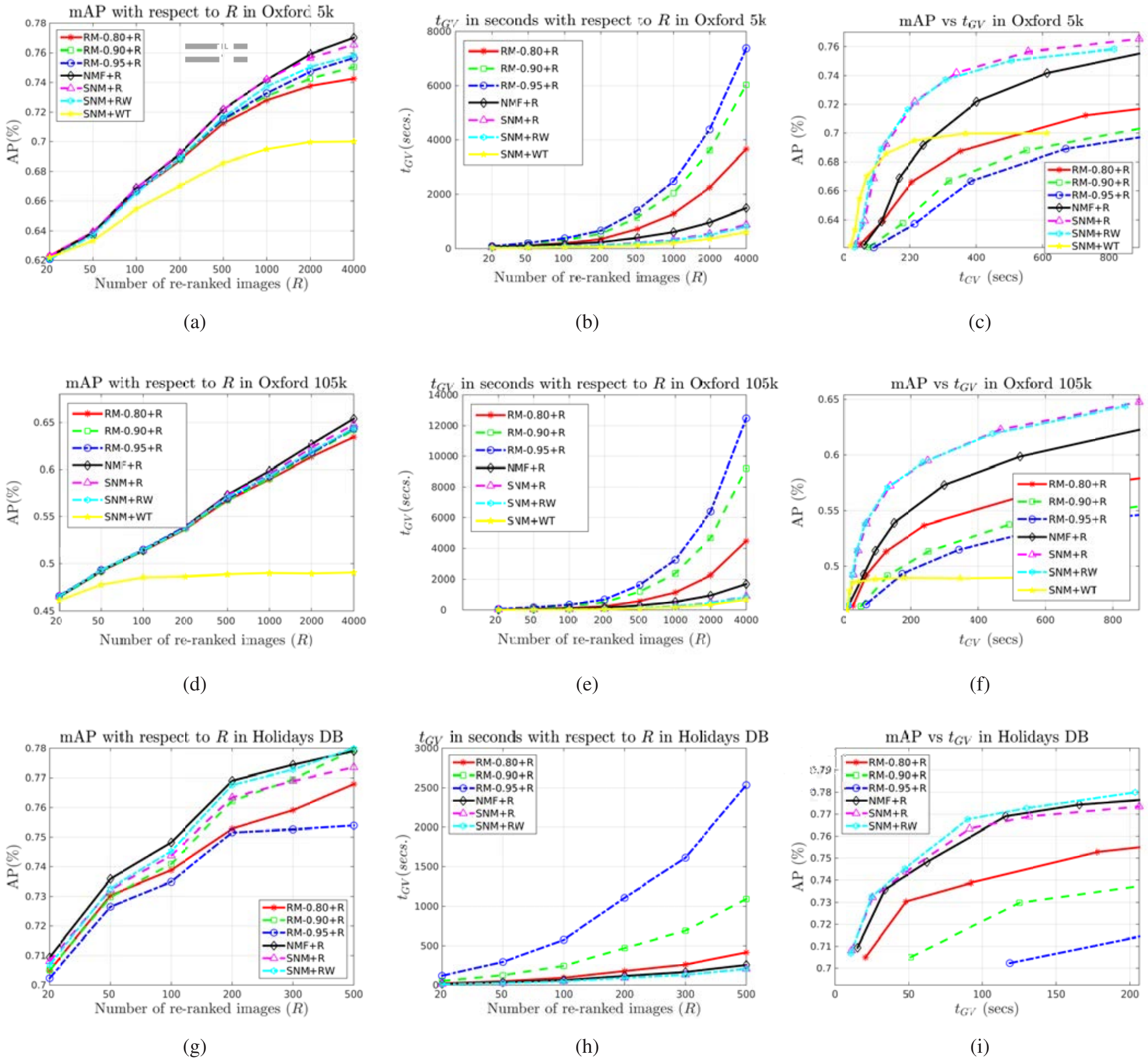


Fig. 4: Results of the use of NM for the geometric verification in image retrieval. Top row: results for the Oxford 5K dataset. Middle row: results for the Oxford 105k dataset. Bottom row: results for the Holidays dataset. Left column: mAP vs. the number of re-ranked images R. Center column: t_{GV} vs. the number of re-ranked images R. Right column: mAP vs. t_{GV}.

associated with the geometrical verification process, we need to estimate first the average execution times by each task of the geometric verification process, namely: the time associated with the computation of matches between every pair of images t_M, the time taken by the NM algorithm t_{NM}, and the time consumed by RANSAC t_R, which is computed from t_h (time to estimate a homography) and t_e (time to evaluate the homography on a particular match). The average values of these variables, which are dataset independent, are shown in Table II.

Next, we need to estimate the average number of keypoints or matches used for each step of the geometric verification process. These values strongly depend on the depth of the re-ranking process R. Let us elaborate a bit more: given that the re-ranking process starts from a baseline ranking provided by a BoW model with a very large vocabulary (which efficiently approximates a matching process between images), it becomes obvious that the number of matches between the query and the ranked list of reference images will decrease as we go down through the ordered list.

The only value that does not depend on R is the average number of keypoints $\bar{N_p}$ detected in the query, which remains constant with R. $\bar{N_p}$ depends both on the keypoint detector and its parametrization, and the visual content of the query. In our experiments, the average number of keypoints detected is $\bar{N_p} = 3,700$ for both Oxford datasets (they share a common list of queries), and $\bar{N_p} = 2,950$ for Holidays.

In Fig. 3 we show the values of $\bar{N_s}$, $\bar{N_g}$ and $\bar{k}$ (described in Section III) as a function of the number of re-ranked images R for the compared algorithms in the Oxford 5k, Oxford 105k and Holidays datasets. In general, we can perceive that the curves are quite similar for all datasets. The main difference between them, which will be analyzed in detail below, comes from the fact that the number of relevant reference images is completely different between datasets: from 1-5 in the Holidays dataset, to a maximum of 220 in the Oxford datasets. This issue affects the behavior on the left part of the curves (in general associated to relevant retrieved images), which might become hidden if only 2-5 images are relevant. Since the Oxford 105k adds distracting images to the Oxford 5k dataset and the depth of the re-ranking step is the same for both datasets (see Section V-D), the behavior in both datasets is very similar.

The average number of strong matches $\bar{N_s}$ resulting from the NM process is equal for all the alternatives and decreases with R, which means that as we go down through the BoW-based ranked list, we encounter images that are more different than the query image with the consequent lower number of matches.

The average number of matches per reference image, $\bar{N_g}$, obviously decreases with R in all the cases since as we go down through the ranked list, the similarity between the query and the reference images also decreases. In the Holidays dataset, in except for some initial positions in the ranking, we are comparing the query with non-relevant images, which causes absolute numbers that are much lower than in the Oxford datasets.

Furthermore, the number of matches considered for geometric verification substantially varies across the different approaches: the baseline approaches based on RM pass every match that satisfies the condition above the distance ratio and, therefore, $\bar{N}_g$ simply varies with the selected $TH_{ratio} = 0.80, 0.90, 0.95$. $NMF + R$ uses NM to filter the matches for which the method does not find neighboring keypoints that also agree on that match, thus reducing $\bar{N}_g$. It is interesting to compare $RM - 0.80 + R$ with $NMF + R$ because in both cases we use $TH_{ratio} = 0.8$. As can be observed for the results of the Oxford 5k and 105k datasets, for depths for which there are many relevant images, the NM algorithm might actually increase the number of matches by also including the weak matches around every strong match. When the relevant images are no longer predominant (higher depths in the re-ranking), the NM algorithm filters false matches and the number of surviving matches is actually lower than that provided by RM. In the Holidays dataset, as the proportion of relevant images is always low for almost any depth in the re-ranking, $NMF + R$ is always faster than $RM - 0.80 + R$. Finally, the alternatives of keeping only the strong matches (SNM, Strong Neighbor Matches) provide notable reductions in the number of matches passed to the geometric verification and, consequently, notable reductions in complexity.

With respect to the average number of samples $\bar{k}$ drawn in the RANSAC robust estimation, it is clear that it mainly depends on the proportion of inliers (true matches) in the whole set of candidate matches. This proportion of inliers is higher when the two images (query and reference) are more similar and, consequently, $\bar{k}$ increases as this proportion decreases with R. It is interesting to note that $\bar{k}$ increases as expected with R, but it starts to slightly decline for higher values of R. This declining happens because, as we have just discussed, the average number of matches entering the geometric verification stage decreases with R and, for high values, the proportion of inliers starts to decline because the lower number of total candidates matches (in the denominator of the inliers proportion) starts to compensate for the lower number of true matches. It is also worth noting how the neighborhood matching helps to identify and remove false matches, what leads to an important increase in the proportion of inliers, thus notably reducing the number of iterations needed in the RANSAC iterative estimation. This capability is indeed improved when combining NM with the weighted version of RANSAC (SNM+WR), which brings out how these weights help RANSAC to choose samples that are free of outliers. Furthermore, the SNM+WT approach substitutes RANSAC by a weighted computation of the projective transformation between images so that $\bar{k} = 1$.

C. Performance-Complexity Analysis of NM for Geometrical Verification

In this section we assess the performance of the proposed alternatives that incorporate NM into the process of geometric verification for reducing its computational complexity. A complete description of the alternatives was given in Section IV-B.

Detailed results for the datasets (Oxford 5k, Oxford 105k, Holidays) are shown in Fig. 4. The rows show the results achieved for the Oxford 5k, Oxford 105k and Holidays datasets, respectively. The first column shows the mean Average Precision (mAP) as a function of the number of re-ranked images R; the second illustrates the evolution of the computational complexity with R; and the third shows the mAP achieved by every method as a function of the computational complexity.

The first two columns, both the performance (mAP) and the complexity (t_{GV}) increase with the number of re-ranked images R. Although some conclusions can be drawn from them, we will focus on the figures of the third column because they provide a fair comparison (for the same complexity level) of the approaches. In general, the proposed NM-based techniques outperform the baseline approaches using RM. In fact, each one of the solutions that combines NM and RANSAC clearly yields better results than the corresponding baseline.

It is also worth noticing how the results varies according to the dataset. For the Oxford datasets, the $SNM + R$ and $SNM + WR$ approaches (those that combine the selection of strong matches with RANSAC) are the best choices because in both cases the complexity of the RANSAC process is dramatically reduced with negligible performance loss. Furthermore, the improvement achieved with respect to their corresponding baselines is consistent for every computational complexity level ($4 - 5\%$ with respect to $RM - 0.8 + R$). Furthermore, in scenarios where very low computational complexity is required, the $SNM + WT$ approach, which substitutes the RANSAC by a direct weighted estimation of the homography based on the strong matches, obtains good performances at very competitive execution times. This result is more notable in the Oxford 5k dataset. As the proportion of relevant images in the top positions of the initial ranking is higher in this dataset than in Oxford 105k, we can conclude that the $SNM + WT$ successfully deals with images in which the proportion of inliers in the candidate set of matches is high, whereas fails to filter out false matches when the number of outliers in the candidate set increases.

For the Holidays dataset, although again the NM-based approaches clearly outperform the corresponding baselines, the results are slightly different. In particular, all solutions combining NM and RANSAC provide comparable performances. On the other hand, $SNM + WT$ provides poor results and we have decided to remove them from the graphs to improve general visualization.

The rationale behind the differences between the datasets is the following. The approaches relying on strong matches (SNM) successfully reduce the number of matches that are passed to RANSAC on those images that show a large number of candidate matches. This occurs in reference images that are

TABLE III: Comparison of several state-of-the-art approaches with the proposed approach for the Oxford 5k and Oxford 105k datasets with respect to mean Average Precision (mAP). In all cases, re-ranking with Geometric Verification has been done using the 1000 top-ranked images. All methods use the vocabularies of BoW in the Paris 6k dataset in except for Mikulík *et al.* $\star$, which used its own larger dataset. For Jegou *et al.*, our own implementation is used †. Some methods, marked as $\diamond$, are not based on BoW and use CNN-based features.

Algorithm	Model	SR	QE	Oxford 5k	Oxford 105k
Philbin *et al.* [25]	BoW	RM	✗	73.1	62.0
Perdǒch *et al.* [52]	BoW	RM	✗	72.5	65.2
Li *et al.* [42]	BoW	RM	✗	73.7	-
Jegou *et al.*† [26]	BoW	RM	✗	77.5	75.0
Proposal ([26]+NM)	BoW	NM	✗	**79.6**	**76.9**
Perdǒch *et al.* [52]	BoW	RM	✓	82.2	77.2
Chum *et al.* [53]	BoW	RM	✓	82.7	76.7
Mikulík *et al.*$\star$ [54]	BoW	RM	✓	84.9	79.5
Arandjelovic *et al.* [55]	BoW	RM	✓	80.9	72.2
Jegou *et al.* † [26]	BoW	RM	✓	83.5	79.6
Proposal ([26]+NM)	BoW	NM	✓	**85.1**	**80.8**
Gordo *et al.*$\diamond$ [18]	CNN	–	✗	84.5	81.6
Gordo *et al.*$\diamond$ [18]	CNN	–	✓	**89.1**	**87.3**

relevant to the query, which is normally more likely in the first positions of the initial ranking provided by the BoW approach. For the Oxford datasets, since hundreds of images are relevant to each query, the improvement provided by these methods is quite notable, diminishing only for high levels of complexity (as can be seen in Fig. 4c, for high values of t_{GV} the curve of $FNM-R$ tends to reach $SNM+R$ and $SNM+RW$). In contrast, for the Holidays dataset, since just a few reference images are relevant to each query, the improvement is related to low levels of complexity (in this case, see Fig. 4i, one can notice that just for low values of t_{GV} the $SNM-R$ and $SNM-RW$ curves are above that of FNM-R).

D. Comparison with the state-of-the-art

To construct an image retrieval system, competitive with the state-of-the-art, we have combined the $NMS+R$ version of our proposal with other modern techniques that have shown to be relevant for the

TABLE IV: Comparison of several state-of-the-art approaches with the proposed approach using the Holidays dataset. In all cases, query expansion methods have been disabled. All methods have used the vocabularies of BoW in the Flickr60K dataset in except for Mikulík *et al.* ⋆, which used its own larger dataset. For Jegou *et al.*, our own implementation is used †. Some methods, marked as ⋄, are not based on BoW and use CNN-based features.

Algorithm	Feat.	SR	Holidays (mAP)
Mikulík et al. ⋆ [54]	BoW	RM	75.8
Li et al. [42]	BoW	RM	**89.2**
Jegou *et al.* † [26]	BoW	RM	82.1
Proposal ([26]+NM)	BoW	NM	83.7
Gordo *et al.*⋄ [18]	CNN	–	86.7
Xie *et al.* ⋄ [56]	CNN	–	**88.7**

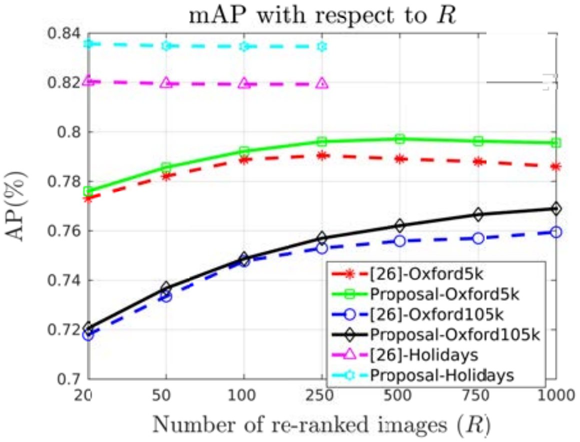


Fig. 5: Results for [26] and our approach at different depths R of the re-ranking step and the three considered datasets.

complete system performance. In general, as our method focuses on the geometric verification step, it can be easily combined with many other techniques described in the image retrieval literature, not only for the initial ranking but also for posterior stages (e.g. Query Expansion). Specifically, in this paper, we have used the following pipeline to generate the initial ranking:

- For the Oxford datasets, we have used the Perdǒch *et al.* [52] detector, which has demonstrated superior performance when all images are vertically aligned.

- We have used *Hamming Embedding (HE)* and *Weak Geometric Constraints (WGC)* [9] rather than the baseline BoW. Hamming embedding was used for a vocabulary of 65k words and binary descriptors of 64 dimensions. Both the vocabulary and the Hamming embedding parameters were trained on the independent datasets (those used in the original proposals): *Paris dataset* [25] was used to compute the vocabulary utilized in the experiments related to the *Oxford datasets*; whereas the *Flickr60K dataset* [9] was used for *Holidays*. Although better results can be obtained if we use the same corresponding dataset to compute the vocabulary (see [55] for some examples), this approach would lead to a non-realistic scenario, prone to over-fitting unless the vocabulary is recomputed every time the reference dataset is increased with new images (which is not scalable at all).
- We have followed the *Multiple Assignment* approach [25], in which each descriptor may be associated to more than one visual word (although, due to the increment of the computational load, we have used this multiple assignment only for the query).
- In order to deal with with the *burstiness phenomenon*, we have also incorporated the model proposed in [26].

We have also made a change regarding the NM-based re-ranking. The NM step of the proposed geometric verification uses *Hamming distances between binary codes* (from Hamming embedding) rather than L_2 distances between SIFT descriptors. In particular, two descriptors are matched only if they are associated with the same word and the hamming distance is small enough. This new approach, although entails a slight decrement on the retrieval performance, turns out to be much more computationally efficient since we only need to compute Hamming distances among a very reduced set of binary descriptors (those that belong to the same word).

Table III shows a comparison between our method and some state-of-the-art techniques which reported results for the Oxford 5k and 105k datasets using the same experimental setup. Two configurations have been tested, with and without Query Expansion (QE), to demonstrate that our approach combines well with QE methods. In particular, we used Discriminative Query Expansion (DQE) [55] since working with more recent alternatives for Query Expansion ([57] or [58]) fall out of the scope of this paper.

We have included various reference methods based on the BoW for comparison, namely: a) the original combination of BoW + geometric verification (Philbin *et al.* [25]); b) discretized local geometry representations (Perdǒch *et al.* [52]); c) fine partitioning of the descriptor space and visual word similarity based on probabilistic relationships (Mikulík *et al.* [54]); d) Average Query Expansion (Chum *et al.* [53]); e) Hamming Embedding with Burstiness and Weak Geometric Constraints (Jegou *et al.* [26]); f) Spatial database-side feature augmentation (SPAUG) and Discriminative Query Expansion (DQE) (Arandjelovic *et al.* [55]); and g) Pairwise Geometric Matching (Li *et al.* [42]). In addition, a couple of methods that are not based on the BoW but use features from CNNs (Xie *et al.* [56], Gordo *et al.*[18]) have been also added at the end of the tables for comparison.

As shown in the tables, our method combines well with other techniques in the literature. In particular, our proposal achieves better performance than all the rest of the approaches based on the BoW paradigm. The results are consistent for the Oxford 5k dataset and the much larger Oxford 105k, which demonstrates its applicability to large-scale image retrieval problems. In fact, as our method is only used for the top images in the initial ranking, its complexity does not directly depend on the size of the dataset, but on the number of re-ranked images. It is also worth mentioning that our method also outperforms [26], being the use of Neighborhood Matching in the re-ranking step the only difference between both methods. Furthermore, our approach would be faster because NM is faster than RM, as we have demonstrated in Section V-C. This particular comparison was extended in Fig. 5 to show results for various depths of the re-ranking process. As can be seen, our method consistently outperforms [26] for any depth. It should be noticed that, in the Oxford 5k dataset, our proposal improves until a depth of around 500 images; while, in Oxford 105k, the performance keeps improving with larger depths. This is due to the quality of the initial ranking provided by BoW, which is lower for the Oxford 105K dataset due to the inclusion of distracting images.

We would also like to remark that the improvements achieved by our method add to those of the QE step. The generation of neighborhoods improves the selection of the final matches and enhances the detection of the ROI (Region Of Interest), factors of key importance to select the most appropriate words and samples in the Query Expansion.

Finally, the performance of methods using CNNs (marked with $\diamond$ in the tables) is slightly better than that achieved by our approach. Although these methods require a previous training phase using large sets of images, the image representations computed by CNNs seem to be more reliable and discriminative than the handcrafted local features used in the traditional BoW approaches.

Results for the Holidays dataset are shown in Table IV. In this case, QE methods have been disabled because they do not achieve improvements due to the small number of relevant images (1-5). For this particular dataset, we have found that the effect of the spatial re-ranking is less relevant for several reasons: lack of planar surfaces (as the building facades in Oxford datasets), lack of initial matches between images, and apparition of objects at different depths (some of these issues will be analyzed in the following section). Consequently, the performance of the proposed method is not so good as those of other methods which focus on other modules of the retrieval pipeline (as [42]) or CNN-based approaches [18][56]. In addition, as can be seen in Fig. 5, the best performance for this dataset is achieved at very low depths of the re-ranking stage, mainly due to the reduced number of relevant images for each query.

Our final conclusion is that NM improves the performance of the re-ranking step in terms of quality and complexity if it is incorporated into a retrieval pipeline based on the BoW paradigm.

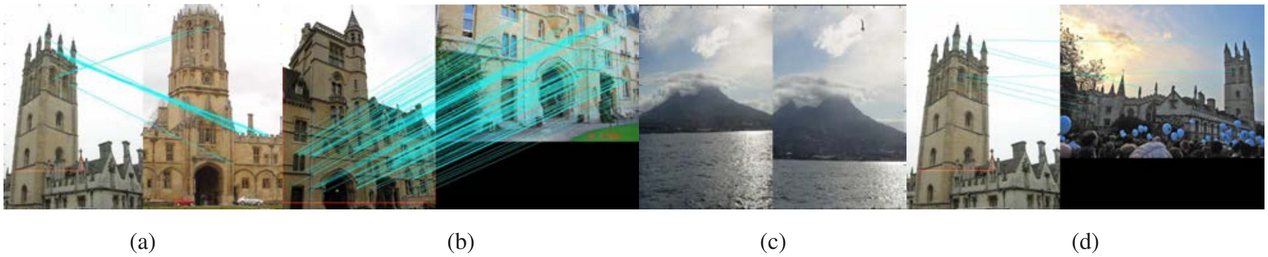
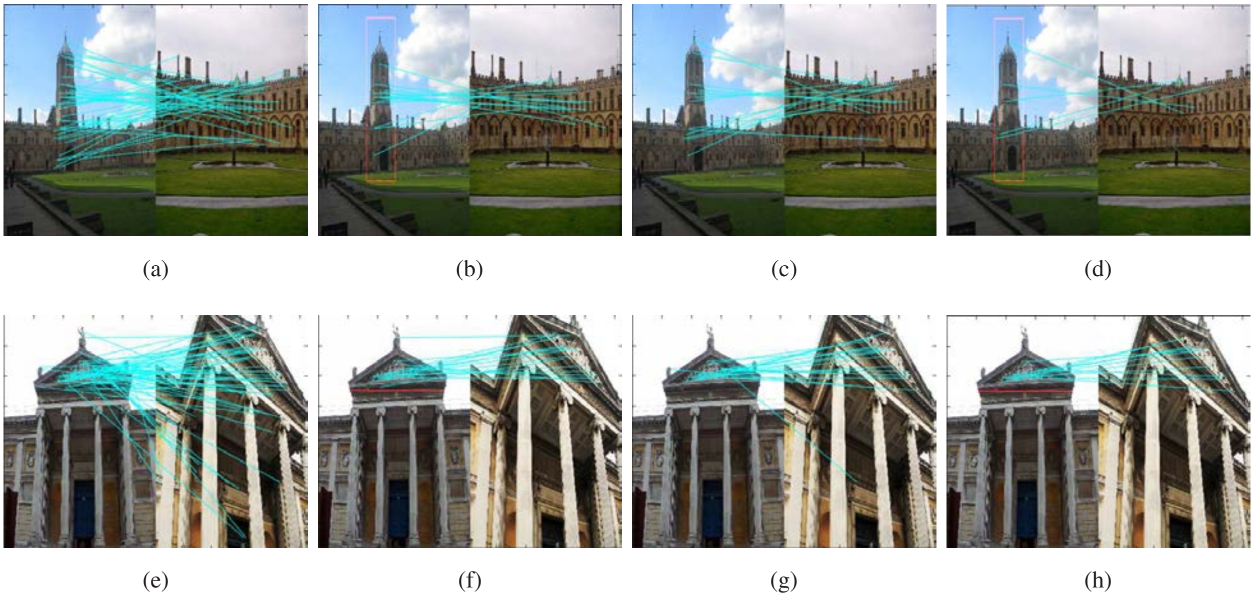


Fig. 6: Examples of errors made by our approach. Non-relevant images retrieved in top positions of the ranking are usually related to buildings that share small details or blocks with the query (a), or correspond with close-ups or partial views of the query (b). Relevant images located deep in the ranking are due to the lack of initial matches (c) and concurrent strong changes in viewpoint, scale and illumination (d).

E. Error Analysis and Discussion

In order to provide more insight about the behavior of our proposed approach, we have first identified cases in which it fails. We have used a version without QE, so that the NM-based re-ranking is the last step in the retrieval pipeline. As we illustrate in Fig. 6, most of the errors cannot be directly related to the Neighborhood Matching process. In fact, non relevant images appear in top positions of the ranking (decrease precision) mainly due to this two reasons: either details and small structures that are shared by several different buildings or landmarks (see elements in the ceilings and towers in Fig. 6 (a)); or images that are close-ups or partial views of larger buildings but have been labeled as negatives (Fig. 6 (b)). Relevant images falling into deep positions of the ranking are more common in the Holidays dataset, due to the small number of relevant images for a query. In some cases, very smooth and homogeneous images causes the previous matching process to remove almost every correspondence and prevents the computing of the geometric transform (Fig 6 (c)). Furthermore, concurrent transformations in geometry, illumination and scale, as shown in Fig. 6 (d) may also break our solution.

Since the most common error sources are not related to the proposed Neighborhood Matching, we have further analyzed the behavior of our NM-based approach (SNM-R version) compared to Regular



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7: Examples of the behavior of our approach. Top row: an example in which our NM-based approach successfully deals with a problematic non-relevant image. The initial set of strong points is $N_s = 43$, containing both true and false matches (a). In regular matching, the geometric verification produces a final set with 13 inliers, which corresponds to the position #20 in the ranking (b). Our approach filters the strong matches that do not belong to a neighborhood resulting in a reduced candidate set with $Nr = 14$ matches (c). The geometric verification results in 5 inliers (d), which corresponds to the position #71 in the ranking. Bottom row: an example in which our approach fails for a relevant image. The initial set of strong matches contains $N_s = 59$ matches (e). The Regular Matching uses these matches to compute the geometric transformation, resulting in a final set of 15 inliers (f) and ranking the image in the position #17. However, our NM-based approach discards strong matches that do not belong to a neighborhood and reduces the candidate set to $N_r = 25$ matches (g). After the geometric verification, our approach finally produces 10 inliers (h) and ranks the image at position #64. This final result shows that the NM stage has removed too many matches, some of them being true.

Matching. In Fig. 7, we show two examples, one in which NM succeeds and other in which fails. The top row shows an example in which our NM-based approach successfully deals with a non-relevant image that is problematic for RM. In this particular case, the NM filtering stage removes many false matches which do not belong to any neighborhood, thus decreasing the final number of geometrically verified matches from 13 (RM) to 5 (NM). In contrast, the bottom row shows an example in which our method fails to process a relevant image. In this case, NM removes too many matches, some of them true, causing the final number of geometrically verified inliers to decrease from 15 (RM) to 10 (NM). However, the first example occurs much more often than the second, which makes that our NM-based approach notably outperforms traditional methods of geometric verification (as we have demonstrated in previous subsections).

VI. Conclusions

In this paper we have described the technique we call *neighborhood matching* (NM) and demonstrated its use for image retrieval. We have discussed the efficacy of NM in reducing the computational complexity of the geometrical verification process employed in image retrieval. The NM technique relies on the hypothesis that if two points are spatially close to each other in an image, it is very unlikely that their corresponding matches in another image are far away from each other. Consequently, this hypothesis is used to filter those likely false matches.

We have proposed several alternative approaches to take advantage of the benefits of NM in image retrieval. Specifically, we have designed various strategies for using NM to reduce the number of potentially false matches that are passed to RANSAC in the geometrical verification step.

We have addressed different image retrieval tasks using the Oxford 5k, Oxford 105k and INRIA Holidays datasets and shown how the use of NM leads to a much better performance-complexity compromise, producing improvements about 5% in terms of mAP with respect to the baseline approach based on RM for equivalent complexity levels. Furthermore, we have compared an improved version of the proposed method with several state-of-the-art methods showing that our method outperforms other methods based on the BoW paradigm in the Oxford 5k and 105k datasets, for which the impact of the geometric re-ranking is higher. As approaches using features computed by CNNs show better performance than those of methods based on the BoW paradigm, our future work will focus on the development of efficient spatial verification methods that can be combined with CNN-based approaches.

References

[1] "Instagram press page," 2015. [Online]. Available: https://instagram.com/press/accessedon21.09.2015

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[3] W. Li, C. Wang, L. Zhang, Y. Rui, and B. Zhang, "Partial-duplicate clustering and visual pattern discovery on web scale image database," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 967–980, July 2015.

[4] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 636–647, May 2015.

[5] C. L. Chou, H. T. Chen, and S. Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 382–395, March 2015.

[6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477. [Online]. Available: http://www.robots.ox.ac.uk/~vgg

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1–8, 2007.

[9] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision (ECCV)*, 2008.

[10] W. Zhou, M. Yang, H. Li, X. Wang, Y. Lin, and Q. Tian, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 601–611, April 2014.

[11] Y. G. Jiang, J. Wang, X. Xue, and S. F. Chang, "Query-adaptive image search with hash codes," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 442–453, Feb 2013.

[12] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 817–824. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995432

[13] H. Xie, K. Gao, Y. Zhang, S. Tang, J. Li, and Y. Liu, "Efficient feature detection and effective post-verification for large scale near-duplicate image search," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1319–1332, Dec 2011.

[14] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 653–661, Aug 2011.

[15] M. Birinci, F. Diaz-de Maria, G. Abdollahian, E. Delp, and M. Gabbouj, "Neighborhood matching for object recognition algorithms based on local image features," in *2011 IEEE Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE)*, Jan 2011, pp. 157–162.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 512–519. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2014.131

[17] L. Zheng, S. Wang, F. He, and Q. Tian, "Seeing the big picture: Deep embedding with contextual evidences," *CoRR*, vol. abs/1406.0132, 2014. [Online]. Available: http://arxiv.org/abs/1406.0132

[18] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," *CoRR*, vol. abs/1604.01325, 2016. [Online]. Available: http://arxiv.org/abs/1604.01325

[19] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008, similarity Matching in Computer Vision and Multimedia.

[20] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Computer Vision ECCV 2002*, ser. Lecture Notes in Computer Science, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Springer Berlin Heidelberg, 2002, vol. 2350, pp. 128–142.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[22] H. Jegou, M. Douze, and C. Schmid, “Improving bag-of-features for large scale image search,” *International Journal of Computer Vision*, vol. 87, pp. 316–336, 2010.

[23] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[24] S. Wei, D. Xu, X. Li, and Y. Zhao, “Joint optimization toward effective and efficient image search,” *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 2216–2227, Dec 2013.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1 –8.

[26] H. Jegou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, June 2009, pp. 1169–1176.

[27] C. H. Lampert, “Detecting objects in large image collections and videos by efficient subimage retrieval.” in *ICCV*. IEEE, 2009, pp. 987–994.

[28] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169–2178.

[29] Y. Ren, A. Bugeau, and J. Benois-Pineau, “Bag-of-bags of words irregular graph pyramids vs spatial pyramid matching for image retrieval,” in *Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on*, Oct 2014, pp. 1–6.

[30] T. Yamasaki and T. Chen, “Spatial statistics for spatial pyramid matching based image recognition,” in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, Dec 2012, pp. 1–10.

[31] K. Kristo and C. S. Chua, “Image representation for object recognition: Utilizing overlapping windows in spatial pyramid matching,” in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3354–3357.

[32] O. Duchenne, A. Joulin, and J. Ponce, “A graph-matching kernel for object categorization,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 1792–1799.

[33] J. Gibert, E. Valveny, and H. Bunke, “Graph embedding in vector spaces by node attribute statistics,” *Pattern Recognition*, vol. 45, no. 9, pp. 3072 – 3083, 2012, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA’2011).

[34] H. Bunke and K. Riesen, “Towards the unification of structural and statistical pattern recognition,” *Pattern Recognition Letters*, vol. 33, no. 7, pp. 811 – 825, 2012, special Issue on Awards from {ICPR} 2010.

[35] I. Gonzalez-Diaz, C. Baz-Hormigos, and F. Diaz-de Maria, “A generative model for concurrent image retrieval and roi segmentation,” *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 169–183, Jan 2014.

[36] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, pp. 381–395, June 1981.

[37] O. Chum, J. Matas, and S. Obdrzalek, “Enhancing ransac by generalized model optimization,” in *Proc. of the ACCV*, vol. 2, 2004, pp. 812–817.

[38] O. Chum and J. Matas, “Optimal randomized ransac,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1472–1482, 2008.

[39] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, “Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, June 2012, pp. 3013–3020.

[40] Y. Zhang, Z. Jia, and T. Chen, “Image retrieval with geometry-preserving visual phrases.” in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 809–816.

[41] Z. Wu, Q. Ke, M. Isard, and J. Sun, “Bundling features for large scale partial-duplicate web image search,” in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.* IEEE, Jun. 2009, pp. 25–32.

[42] X. Li, M. Larson, and A. Hanjalic, “Pairwise geometric matching for large-scale object retrieval,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, June 2015, pp. 5153–5161.

[43] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, “Descriptive visual words and visual phrases for image applications,” in *Proceedings of the 17th ACM International Conference on Multimedia*, ser. MM ’09. New York, NY, USA: ACM, 2009, pp. 75–84. [Online]. Available: http://doi.acm.org/10.1145/1631272.1631285

[44] W. Zhou, H. Li, Y. Lu, and Q. Tian, “Sift match verification by geometric coding for large-scale partial-duplicate web image search,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 1, pp. 4:1–4:18, Feb. 2013. [Online]. Available: http://doi.acm.org/10.1145/2422956.2422960

[45] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian, “Toward a higher-level visual representation for object-based image retrieval,” *The Visual Computer*, vol. 25, no. 1, pp. 13–23, 2009. [Online]. Available: http://dx.doi.org/10.1007/s00371-008-0294-0

[46] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[47] M. Muja and D. G. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *International Conference on Computer Vision Theory and Application VISSAPP’09)*, 2009.

[48] J. Philbin and A. Zisserman, “Oxford building dataset,” Website, http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/.

[49] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *Int. J. Comput. Vision*, vol. 60, pp. 63–86, October 2004.

[50] E. Zhang and Y. Zhang, *Average Precision*. Boston, MA: Springer US, 2009, pp. 192–193. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-39940-9_482

[51] M. Muja and D. G. Lowe, “Scalable nearest neighbor algorithms for high dimensional data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, 2014.

[52] M. Perďoch, O. Chum, and J. Matas, “Efficient representation of local geometry for large scale object retrieval,” in *CVPR 2009: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Madison, USA: Omnipress, June 2009, pp. 9–16.

[53] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, “Total recall ii: Query expansion revisited,” in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 889–896. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995601

[54] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, “Learning a fine vocabulary,” in *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ser. ECCV’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–14.

[55] R. Arandjelovic, “Three things everyone should know to improve object retrieval,” in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 2911–2918. [Online]. Available: http://dl.acm.org/citation.cfm?id=2354409.2355123

[56] L. Xie, R. Hong, B. Zhang, and Q. Tian, “Image classification and retrieval are one,” in *Proceedings of the 5th ACM*

on International Conference on Multimedia Retrieval, ser. ICMR ’15. New York, NY, USA: ACM, 2015, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/2671188.2749289

[57] L. Xie, Q. Tian, W. Zhou, and B. Zhang, “Heterogeneous graph propagation for large-scale web image search,” *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 4287–4298, 2015. [Online]. Available: http://dx.doi.org/10.1109/TIP.2015.2432673

[58] Z. Zhong, J. Zhu, and S. C. H. Hoi, “Fast object retrieval using direct spatial matching,” *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1391–1397, Aug 2015.

Iván González-Díaz Iván González-Díaz received the Telecommunications Engineering degree from Universidad de Valladolid, Valladolid, Spain, in 1999, the M.Sc. and Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007 and 2011, respectively. After holding a postdoc position in the Laboratoire Bordelais de Recherche en Informatique at the University Bordeaux, he currently works as a Visiting Lecturer at the Signal Theory and Communications Department in Universidad Carlos III de Madrid. His primary research interests include object recognition, category-based image segmentation, scene understanding and content-based image and video retrieval systems. In these fields, he is co-author of several papers in prestigious international journals, two chapters in international books and a few papers in revised international conferences.

Murat Birinci Murat Birinci has received his BSc degree in 2005 in Electrical and Electronics Engineering from Middle East Technical University (Turkey), and his MSc degree in 2007 in Image and Video Signal Processing from Tampere University of Technology (Finland) where he is currently pursuing his PhD degree in the same field. He has also worked in Nokia Research Center and currently employed by Intel working on camera control algorithms. His research interests are visual perception, image and video analysis, image segmentation, object detection and recognition.

Fernando Díaz-de-María Fernando Daz-de-Mara (M’97) received the Telecommunication Engineering degree and the Ph.D. degree from the Universidad Politcnica de Madrid, Madrid, Spain, in 1991 and 1996, respectively. Since October 1996, he has been an Associate Professor in the Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Madrid, Spain.

His primary research interests include video coding, image and video analysis, and computer vision. He has led numerous projects and contracts in the fields mentioned. He is co-author of numerous papers in peer-reviewed international journals, several book chapters and a number of papers in national and international conferences.

Edward J. Delp (S'70,M'79,SM'86,F'97,LF'14) was born in Cincinnati, OH. He received the B.S.E.E. (cum laude) and M.S. degrees from the University of Cincinnati and the Ph.D. degree from Purdue University, West Lafayette, IN. In May 2002, he received an Honorary Doctor of Technology from Tampere University of Technology, Tampere, Finland.

From 1980 to 1984, he was with the Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor. Since August 1984, he has been with the School of Electrical and Computer Engineering and the School of Biomedical Engineering, Purdue University. In 2008 he was named a Distinguished Professor and is currently The Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering. His research interests include image and video compression, multimedia security, computer vision, medical imaging, multimedia systems, communication, and information theory.

Dr. Delp is a Fellow of the SPIE, a Fellow of the Society for Imaging Science and Technology (IS&T), and a Fellow of the American Institute of Medical and Biological Engineering. In 2004 he received the Technical Achievement Award from the IEEE Signal Processing Society for his work in image and video compression and multimedia security. In 2008 Dr. Delp received the Society Award from the IEEE Signal Processing Society (SPS). This is the highest award given by SPS and it cited his work in multimedia security and image and video compression.